INF 558 – Spring 2019

Homework #7: String and Data Matching

Due: 4/17, Wednesday
100 points
(TA handling this homework: Binh Vu binhlvu@usc.edu)

Task 1 – String similarity (60 pts)

In this homework, you are given two data sets **fodors.csv** and **zagats.csv** which contains information about restaurants from 2 different sources. Each dataset contains 3 different fields: name, phone and cuisine.

Your goal in this homework is to match records from these 2 datasets using entity linking methods. This means you need figure out which pairs of records (restaurants) in the 2 datasets are referring to the same restaurant. (Hint: there are 112 matches between the 2 datasets)

Analyze the given data and choose string similarities that you think are appropriate for each field. Explain your choices in the report.

Write a program that computes the field similarities between records from 2 datasets.

Task 2 – Data/Record Matching (30 pts)

Design a scoring function to combine your field similarities. Explain your choices of weights in the scoring function in the report.
Write a program to compute the overall record similarity based on your field similarities and export an output file with the following format:

[filename of source1:record's id][TAB character][filename of source2:record's id]

For example:
zagat.csv:49 fodors.csv:358
zagat.csv:64 fodors.csv:297

This means records #49 and #64 in zagats.csv match with records #358 and #297 in fodors.csv, respectively.

The ground truth of record matching will not be released. Your output file will be graded based on recall, precision and F-measure after your submission.

Task 3 – Scaling up (10 pts)

Implement an approach to reducing the number of comparisons you need to perform. Report the number of comparisons the improved approach performs. Does it miss any matching records? If yes, provide an example. If not, explain why.

Submission Instructions

You must submit the following files a single .zip archive named Firstname_Lastname_hw7.zip and submit it via Blackboard:

- Firstname_Lastname_hw7_report.pdf: A pdf file containing answers for Task 1, 2 and 3
- src: A folder which contains all the source code you wrote for this homework with a README file on how to run and export your output file.
- output.txt: A txt file which contains your output.